CLAIMS

What is claimed is:

1. A method comprising:

identifying a first set of anchor text written in a first format and containing a given term;

identifying a set of documents to which the first set of anchor text points;

identifying a second set of anchor text written in a second format and pointing to the identified set of documents;

analyzing the second set of anchor text to determine that a representation of the given term in the first format corresponds to a representation of the given term in the second format.

2. The method of claim 1, in which the first format comprises a first character set, and the second format comprises a second character set.

3. The method of claim 1, in which the first format comprises a first language and the second format comprises a second language.

4. The method of claim 1, in which analyzing the second set of anchor text includes identifying a term that appears most frequently in the second set of anchor text and designating the most frequently appearing term as the representation of the given term in the second format.

5. The method of claim 1, in which analyzing the second set of anchor text comprises:

calculating a probability that the given term corresponds to a term in the second set of anchor text.

6. The method of claim 5, in which the probability is obtained using at least one of Bayesian methods, histogram smoothing, kernel smoothing, and shrinkage estimators.

7. The method of claim 5, in which the probability that the given term corresponds to a term in the second set of anchor text is obtained by dividing the number of occurrences of the term in the second set of anchor text by the total number of occurrences of all terms in the second set of anchor text.

8. The method of claim 1, in which analyzing the second set of anchor text comprises:

calculating a probability that the given term corresponds to each term in the second set of anchor text.

9. The method of claim 1, in which analyzing the second set of anchor text comprises:

identifying a term that appears most frequently in the second set of anchor text.

10. The method of claim 2, in which the first format is selected from the group consisting of: romaji, romaja, and pinyin; and in which the second character set is selected from the group consisting of: katakana, hiragana, kanji, hangul, hanja, and traditional Chinese characters.

11. The method of claim 1, in which the documents comprise web pages.

12. The method of claim 1, further comprising:

obtaining a query written in the first format and containing the given term;

translating the query into the second format based at least in part on said analyzing step;

searching a database for information written in the second format that is responsive to the translated query.

13. The method of claim 12, in which the steps are performed in the order recited.

14. A search method comprising:

obtaining a query written in a first format from a user;

translating the query into a second format using a probabilistic dictionary, the probabilistic dictionary mapping terms from the first format to the second format;

searching a database for information responsive to the translated query; and

returning search results written in the second format to the user.

15. The search method of claim 14, further comprising:

obtaining search result selections from the user;

using said search result selections to modify the probabilistic dictionary of term mappings.

16. The search method of claim 15, wherein the modification comprises adjusting at least one probability associated with at least one mapping in the probabilistic dictionary.

17. The search method of claim 14, in which the step of translating the query into the second format includes expanding the query.

18. The search method of claim 17, in which the expanded query includes alternative encodings of the query terms.

19. The search method of claim 17, in which the expanded query includes alternative language translations of the query terms.

20. The search method of claim 17, in which the expanded query includes alternative encodings and alternative language translations of the query terms.

21. The search method of claim 18, in which the expanded query includes synonyms of the alternative encodings of the query terms.

22. A method for creating a probabilistic dictionary, the probabilistic dictionary mapping terms in a first format to terms in a second format, the method comprising:

for a given term, identifying a first set of data in the first format that contains the term;

identifying a second set of data in the second format that is aligned with the first set of data; and

analyzing the second set of data to determine one or more probabilities with which the given term maps onto one or more terms in the second set of data.

23. The method of claim 22, further comprising:

adding the given term to the dictionary along with one or more probabilities with which the given term maps onto one or more terms in the second set of data.

24. The method of claim 23, further comprising:

repeating, for each term to be added to the dictionary, said steps of identifying a first set of data, identifying a second set of data, and analyzing the second set of data.

25. The method of claim 22, in which the first set of data comprises a first set of anchor text pointing to a set of one or more web pages, and in which the second set of data comprises a second set of anchor text pointing to the same set of one or more web pages.

26. The method of claim 22, in which the first set of data comprises a set of text written in a first language, and in which the second set of data comprises the same set of text written in a second language.

27. The method of claim 22, in which the probability with which the given term maps onto a term in the second set of data is calculated by dividing the number of occurrences of the term in the second set of data by the total number of terms in the second set of data.

28. The method of claim 22, further comprising:

modifying the probability with which the given term maps onto a term in the second set of data based, at least in part, on an analysis of a user's selection of search results.

29. The method of claim 22, further comprising:

modifying the probability with which the given term maps onto a term in the second set of data based, at least in part, on an analysis of a user's previous queries.

30. A computer program product embodied on a computer-readable medium, the computer program product including instructions, which when executed by a computer system, are operable to cause the computer system to perform acts comprising:

identifying a first set of anchor text written in a first format and containing a given term;

identifying a set of web pages to which the first set of anchor text points;

identifying a second set of anchor text written in a second format and pointing to the identified set of web pages;

determining a probability that a representation of the given term in the first format corresponds to a representation of the given term in the second format.

31. The computer program product of claim 30, further including instructions, which when executed by the computer system, are operable to cause the computer system to perform acts comprising:

modifying the probability that a representation of the given term in the first format corresponds to a representation of the given term in the second format based, at least in part, on an analysis of a user's selection of search results.

32. The computer program product of claim 30, further including instructions, which when executed by the computer system, are operable to cause the computer system to perform acts comprising:

modifying the probability that a representation of the given term in the first format corresponds to a representation of the given term in the second format based, at least in part, on an analysis of a user's previous queries.

33. The computer program product of claim 30, in which the probability is determined, at least in part, using at least one of Bayesian methods, histogram smoothing, kernel smoothing, and shrinkage estimators.

34. A translation method comprising:

identifying a first body of text written in a first format;

identifying a second body of text written in a second format, the second body of text being aligned with the first body of text;

creating a dictionary of translations between terms in the first body of text and terms in the second body of text by comparing the occurrence of terms in the first body of text with the occurrence of terms in the second body of text.

35. A translation method as in claim 34, in which the dictionary of translations includes one or more probabilities associated with the translations.

36. A translation method as in claim 34, in which the first format comprises a first character set and the second format comprises a second character set.

37. A translation method as in claim 34, in which the first format comprises a first language and the second format comprises a second language.

38. A translation method as in claim 34, in which the first body of text comprises anchor text and the second body of text comprises anchor text.

39. A method comprising:

receiving a query containing at least one query term written in a first format;

translating the query term into a plurality of variants written in a second format; and

using one or more of the variants to search for information written in the second format that is responsive to the query.

40. The method of claim 39, in which the first format comprises a sequence of numbers entered from a telephone keypad; and in which the second format comprises alphanumeric text.

41. The method of claim 39, further comprising:

obtaining the one or more variants by discarding variants in the plurality of variants that are not part of a predefined lexicon.

42. The method of claim 39, further comprising:

obtaining the one or more variants by discarding variants in the plurality of variants that contain predefined low-probability character combinations.

43. The method of claim 39, in which the first format comprises alphanumeric text written in a character set selected from the group consisting of romaji, romaja, and pinyin; and in which the second format comprises alphanumeric text written in a character set selected from the group consisting of kanji, katakana, hiragana, hangul, hanja, and traditional Chinese characters.

44. A method comprising:

receiving a numeric query entered from a telephone keypad;

translating the numeric query into a group of potential alphanumeric translations in a first format;

discarding potential translations that are determined to include predefined low-probability character combinations;

translating the remaining alphanumeric translations from the first format to a second format using a probabilistic dictionary; and

performing a search using the alphanumeric translations in the second format.

45. The method of claim 44, in which the first format comprises text written in a character set selected from the group consisting of romaji, romaja, and pinyin; and in which the second format comprises text written in a character set selected from the group consisting of kanji, katakana, hiragana, hangul, hanja, and traditional Chinese characters.